Validation of Bone Age Methods by Their Ability to Predict Adult Height

Thodberg, Hans Henrik ; Neuhof, Julia ; Ranke, Michael B ; Jenni, Oskar G ; Martin, David D

Abstract: AIM: Several bone age (BA) methods are in use today. The aim of this study was to introduce a framework for assessing the validity of a BA method by its ability to predict adult height (H) and to apply it to manual ratings based on Greulich-Pyle (GP) and Tanner-Whitehouse 3 (TW) and to the fully automated BoneXpert method. MATERIAL: The study used X-rays of 232 children from the First Zurich Longitudinal Study recorded close to each anniversary. METHOD: For each height measurement (h), we calculated the growth potential (gp), defined as gp = (H-h)/H. The standard deviation of the gp prediction error for children of the same age was taken as a measure of the validity of the BA method and averaged over the age range 10-18 years for boys and 8-16 years for girls to obtain the overall gp prediction error (GPPE). RESULTS: Manual TW yielded GPPE = 1.32% [95% CI 1.28-1.36], and was significantly outperformed by manual GP with GPPE = 1.26% [1.22-1.30]. The automated rating obtained GPPE = 1.23%, and omitting radius and ulna yielded GPPE = 1.22%. CONCLUSION: Manual GP rating is better than manual TW rating in predicting adult height, and the fully automated method works as well as manual GP rating.

DOI: https://doi.org/10.1159/000313592

Original Paper

HORMONE RESEARCH IN PÆDIATRICS

Horm Res Paediatr 2010;74:15–22
DOI: 10.1159/000313592

Received: May 11, 2009
Accepted: September 23, 2009
Published online: April 20, 2010

Validation of Bone Age Methods by Their Ability to Predict Adult Height

Hans Henrik Thodberg[a] Julia Neuhof[b] Michael B. Ranke[b] Oskar G. Jenni[c] David D. Martin[b]

[a]Visiana, Holte, Denmark; [b]University Children's Hospital Tübingen, Germany; [c]Child Development Center, University Children's Hospital, Zurich, Switzerland

Key Words
Automatic bone age · Greulich-Pyle · BoneXpert · Skeletal maturity · Height prediction · Bayley-Pinneau · Growth potential

Abstract
Aim: Several bone age (BA) methods are in use today. The aim of this study was to introduce a framework for assessing the validity of a BA method by its ability to predict adult height *(H)* and to apply it to manual ratings based on Greulich-Pyle (GP) and Tanner-Whitehouse 3 (TW) and to the fully automated BoneXpert method. ***Material:*** The study used X-rays of 232 children from the First Zurich Longitudinal Study recorded close to each anniversary. ***Method:*** For each height measurement *(h)*, we calculated the growth potential (*gp*), defined as $gp = (H - h)/H$. The standard deviation of the *gp* prediction error for children of the same age was taken as a measure of the validity of the BA method and averaged over the age range 10–18 years for boys and 8–16 years for girls to obtain the overall *gp* prediction error (GPPE). ***Results:*** Manual TW yielded GPPE = 1.32% [95% CI 1.28–1.36], and was significantly outperformed by manual GP with GPPE = 1.26% [1.22–1.30]. The automated rating obtained GPPE = 1.23%, and omitting radius and ulna yielded GPPE = 1.22%. ***Conclusion:*** Manual GP rating is better than manual TW rating in predicting adult height, and the fully automated method works as well as manual GP rating.

Introduction

This study presents a framework for validation of bone age (BA) methods. The motivation for setting up this framework is that there are several 'competing' BA methods which are each defined within their own particular paradigm. The most common BA methods are the manual Tanner-Whitehouse (TW) method [1] and the manual Greulich-Pyle (GP) method [2], and recently the completely automated BoneXpert method was introduced [3, 4]. These methods work by different principles and they each have their merits. For instance the TW method compels the reader to consider 13 bones separately and employs a standardized way of combining information from the bones; the GP method is fast, easy and intuitive, and the BoneXpert method is time-saving and eliminates the rater variability.

The fact that automated ratings are objective does not necessarily mean that their BA values are more clinically

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1663–2818/10/0741–0015$26.00/0

Accessible online at:
www.karger.com/hrp

Dr. David D. Martin
Paediatric Endocrinology Section, University Children's Hospital
Hoppe-Seyler Strasse 1, DE–72076 Tübingen (Germany)
Tel. +49 7071 29 83795, Fax +49 7071 29 4157
E-Mail david.martin@med.uni-tuebingen.de

relevant than manual ratings. The inner workings of BoneXpert is a sophisticated mathematical model that analyses the image in a presumably very different way than the human mind, and one could fear that, for all their reproducibility, automated ratings might not reflect the 'true BA', but rather some slightly different property of the X-ray.

The problem is that there is no direct way to determine the 'true BA', no golden standard. We cannot dissect the bones to find out how mature they are. Instead BA is loosely defined as the age at which a similar Gestalt of the radiograph is observed in healthy children. The word Gestalt alludes to a kind of holistic or intuitive judgment of the appearance of the bones. Humans are remarkably fast at making judgments of Gestalts, but there always remains an element of subjectivity. The various manual BA rating methods attempt to eliminate this rater variability by specifying certain maturity indicators, and by training of the raters, but this can never be completely accomplished. As a result, if a set of hand X-rays is ordered according to maturity by the TW BA system and by the GP BA system, these orderings will in general be different, and it seems difficult – or even meaningless – to decide which is the most valid.

This work proposes to use a BA method's ability to predict adult height as a measure of its validity, i.e. a completely objective validation framework. The prediction of adult height is one of the most common applications of BA assessments, used for instance as an element in the diagnosis of short or tall stature. Admittedly, adult height cannot be predicted exactly from age, BA and current height, and there are also other important applications of BA, so this validation framework is not entirely sufficient, but its simplicity and objectivity makes it scientifically interesting.

Methods

The Heritage of Bayley and Pinneau

The proposed framework is rooted in the work of Bayley and Pinneau [5], who made retrospective studies of healthy children of known BA, height *(h)* and adult height *(H)*. They divided their observations into groups of the same sex and chronological age (CA), and on computing the percentage of mature height (PMH) that each child had attained as

$$\mathrm{PMH} = 100\ h/H,$$

they found a correlation (Pearson's ρ) between PMH and BA of for instance 0.86, implying that BA accounted for $0.86^2 = 74\%$ of the variation of PMH, and they concluded that BA was an important predictor of adult height. They proceeded to construct tables of PMH predictions based on BA and CA, as follows:

$$H_{\mathrm{pred}} = h \times 100/\mathrm{PMH}$$

Unfortunately they produced only one-way tables of PMH versus BA. It would have been more accurate to produce two-way tables of PMH versus any pair of (BA, CA). Instead they divided the cases into three groups according to BA – CA, i.e. BA more than 1 year retarded, normal BA and BA more than 1 year advanced, with corresponding one-way tables. This is perhaps the most serious limitation of their work.

The Validation Framework

In this work we considered instead of PMH an equivalent quantity, referred to as the *growth potential*:

$$gp = (H - h)/H$$

The correlations of BA to *gp* and PMH are exactly the same, since *gp* is the 'mirror image' of PMH: $gp = 1 - \mathrm{PMH}/100$.

While Bayley and Pinneau wanted to predict adult height, our aim was to validate a BA method. We defined a valid BA method as one that gave an accurate prediction of *gp*. This was quantified by the prediction SD of the fit of *gp* versus BA for each CA. This SD was averaged over a range of CAs to produce an overall, robust measure.

Data

The data were the digitized left- and right-hand X-rays of the First Zurich Longitudinal Study (1ZLS). The basic properties of BoneXpert GP BA have been studied previously [3, 6, 7]. The age range was 2–20 years and the 232 included children were examined once a year. 94% of all images were taken within 2 weeks and 99% within 1 month of the respective child's birthday; in this study they were assumed to have been taken exactly on the child's birthday. All children were examined yearly until adult height *(H)*, defined as the height by which the growth was <0.5 cm over the last 2 years.

Bone Age Methods

The following BA methods were validated with the new framework:

(1) *Manual TW BA:* this BA was derived from the original TW stage ratings of the left-hand films performed by several experienced TW raters at the time of the 1ZLS. From the maturity stages of the 13 RUS bones (radius, ulna and the short bones of rays 1, 3 and 5) a BA was computed (many years later) using the TW3 reference [1]. In the CA interval 10–18 years for boys and 8–16 years for girls, 97% of these BA ratings were available.

(2) *Manual GP:* this was the original GP BA rating of the left hand performed at the time of the 1ZLS. To our knowledge, these GP ratings have not been used in any previous publications – the researchers in Zurich seem to have preferred the TW ratings. In the CA interval 10–18 years for boys and 8–16 for girls, 92% of these ratings were present.

(3) *BoneXpert GP:* this was the automated GP BA rating by BoneXpert, available for both the left and the right hands. This BA is based on the same 13 bones as used in the TW RUS system. For both hands, 95% of the ratings were present in the age intervals 10–18 years for boys and 8–16 for girls. These ratings have also been studied by Martin et al. [8].

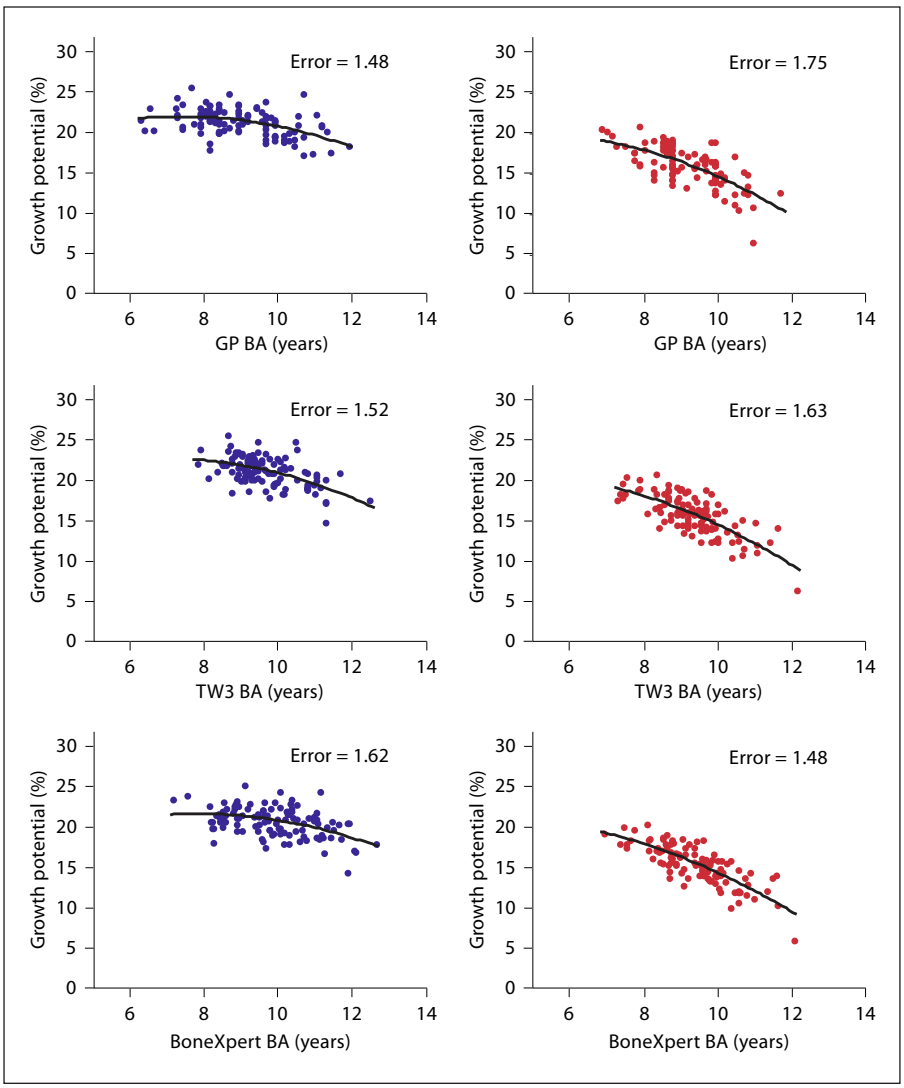


Fig. 1. Growth potential versus manual GP (top), manual TW3 (middle) or BoneXpert (bottom) BA in boys (left) and girls (right) at the CA of 10 years. Explanation: 10-year old girls with BA 8 years had approx. 18% of their adult height left to grow, while those with BA 11 years had only approx. 12% left.

(4) *BoneXpert GPshort:* this was BoneXpert's GP BA rating of the left and the right hand excluding radius and ulna, i.e. using only the 11 short bones in rays 1, 3 and 5. This variant was studied to assess the relevance of including the wrist in the BA rating.

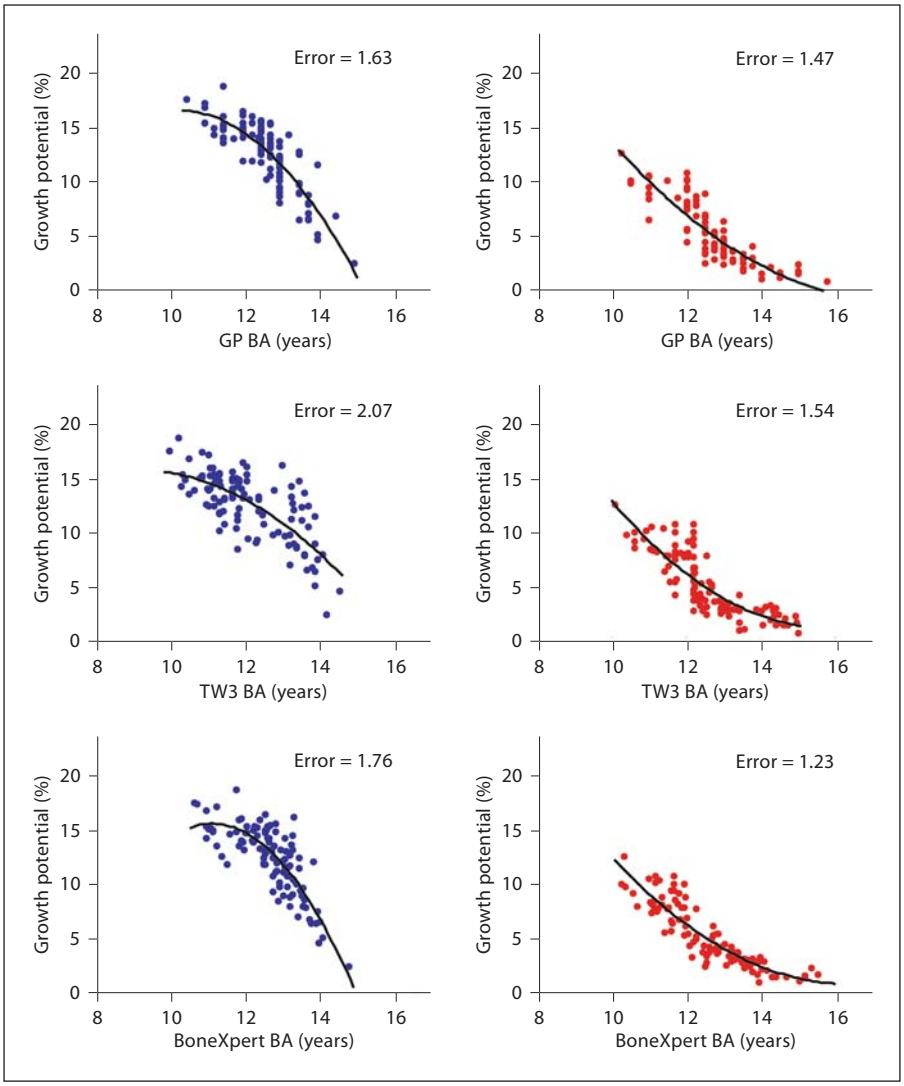
Results

Figures 1–3 show the growth potential *gp* versus the manual TW3, manual GP and automated GP BA at three selected CAs, 10, 13 and 16 years, which represent children before, and in early and late puberty.

At 13 years, the correlation between *gp* and GP BA was 0.82 for boys and 0.85 for girls. However, the relation between *gp* and BA was in general non-linear, so the (linear) correlation is not a faithful representation of the degree of relationship between the two quantities. Therefore, the relation was modeled by a second-order polynomial, i.e. a parabola was fitted to each plot. This can be seen to be an acceptable approximation in figures 1–3. The SD of the fit is indicated in each plot.

The analysis was done at all integer ages from 5 to 19, and the SD errors for the *gp* predictions at these ages are shown in figure 4. It is seen that GP yielded lower errors than TW3, except for girls aged 9–12 years. The automated method outperformed the manual methods in girls but not in boys.

To form an overall performance value, the errors were averaged over the CA intervals 10–18 years for boys and



Color version available online

Fig. 2. Growth potential versus BA at the CA of 13 years. Age 13 was close to the growth spurt for both boys and girls, hence the growth potential depends dramatically on maturity. For the boys, the GP BA described the growth potential much better than TW3 BA, which seemed to be 'confused' around 13 years.

8–16 years for girls. This average SD, termed growth potential prediction error (GPPE), is presented for all five BA methods in table 1 for boys and girls separately, as well as averaged over the two sexes.

For the automated method, the analyses were made twice: once on the left hand and once on the right-hand X-rays. The results were very similar, i.e. the left and the right hand were equally good at predicting adult height. Table 1 lists the average of the SDs obtained in the two hands.

The GPPE values for the two sexes combined are based on 1,914 images for GP BA and 2,013 images for TW BA, yielding the 95% confidence intervals in table 1. Manual GP was significantly better than manual TW3 ($p < 0.05$). BoneXpert GP and BoneXpert GPshort performed virtually the same, and both were better than manual GP, although not significantly better.

Table 1. GPPE for four different BA methods

BA system	Boys	Girls	Both sexes
Manual TW3	1.30	1.33	1.32 [1.28; 1.36] 95%
Manual GP	1.20	1.33	1.26 [1.22; 1.30] 95%
BoneXpert GP	1.29	1.16	1.23 [1.19; 1.27] 95%
BoneXpert GPshort	1.28	1.16	1.22 [1.18; 1.26] 95%

We have tried to use a third-order polynomial to fit all the *gp* curves. This reduced the GPPEs of the method by 0.02 or 0.03, but did not change the relative merit of the

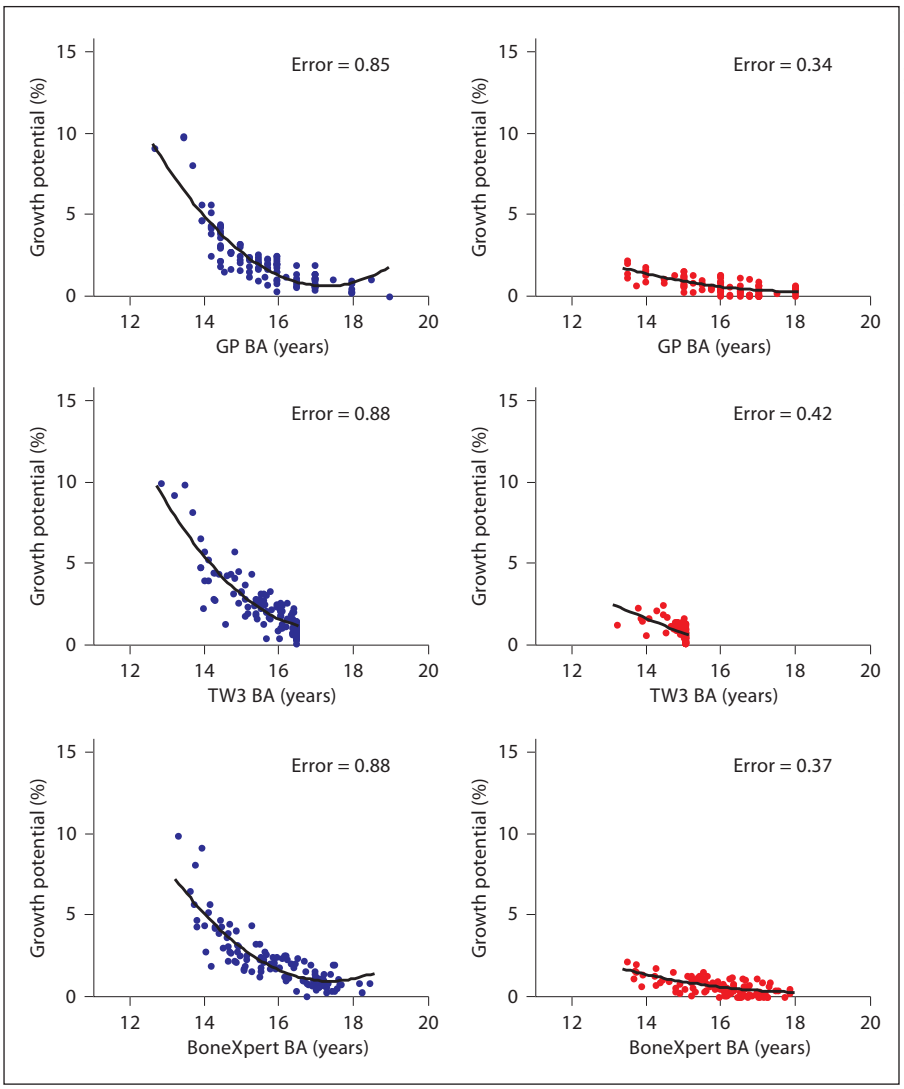


Fig. 3. Growth potential versus BA at the CA of 16 years. At age 16 years, the GP and BoneXpert BAs described the end of the growth potential curve well, while the TW3 BA system terminated at 16.5 years for boys and at 15 years for girls.

different methods. In some cases, for instance in boys of GP 16 years in figure 3, this fit made more sense as it avoided the upturn after BA >16, but then for several other cases the extra degree of freedom lead to spurious swings, which did not make sense, so a quadratic polynomial was the overall best choice. The intention of these fits was not to form a model for height prediction, so these few artifacts do not matter.

Discussion

The Framework

The reader might wonder why these data were not analyzed in terms of a conventional model for adult height prediction, i.e. one that predicts adult height based on current CA, BA and height. The reason is that there are several different models to choose among [9–11]. Furthermore, these models depend on the population of subjects used to estimate them, and on the idiosyncrasies of the BA rater who provided the BA values for the models. One would therefore need to re-estimate these models on the 1ZLS data, and to enter into a discussion of the merits of different models [12].

The proposed framework is much simpler. It is based on the insight that the growth potential *(gp)* of children of the same age is the essential quantity that we want BA to describe. This relation is expected to be different at different ages, because there are different phases of growth, so each age was analyzed separately in order not

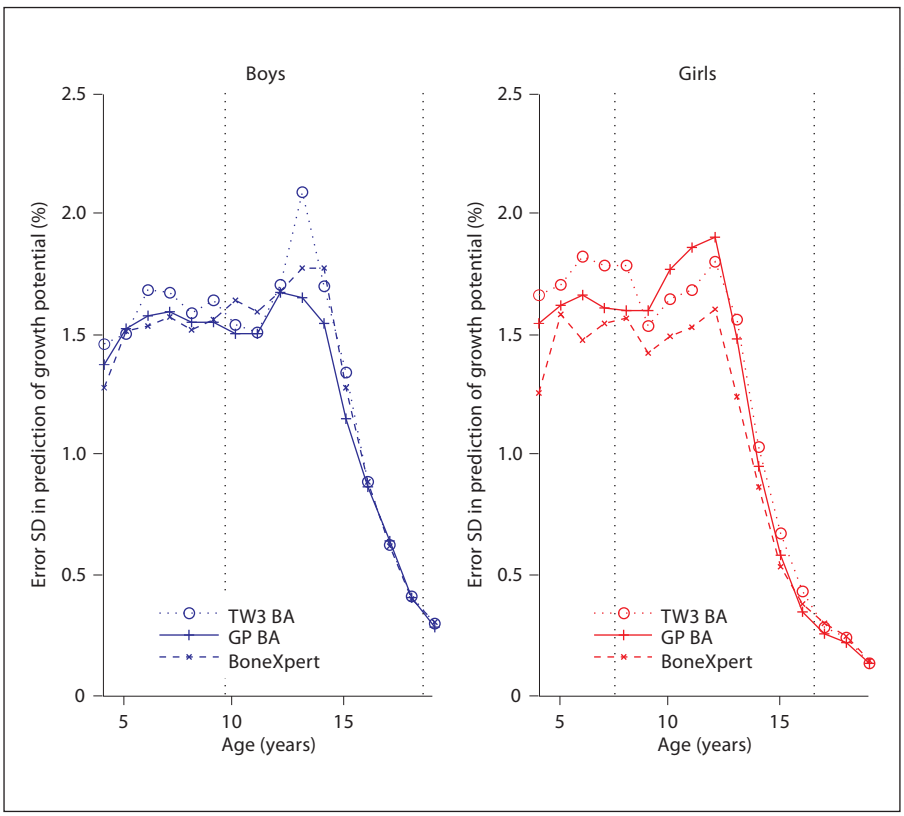


Fig. 4. Validation of the two manual BA rating methods (GP and TW3) and a fully automated method (BoneXpert) on the 1ZLS. GP and BoneXpert significantly outperformed TW3 in prediction of growth potential. The GPPE (table 1) was computed as the average of the error SDs between the vertical, dotted lines.

to assume any specific model. And the functional relationship between BA and *gp* was not assumed to be linear. In fact, figures 1–3 clearly show that it is non-linear (incidentally, this points to a problem in the TW Mark II [11] and RWT [10] methods for adult height prediction, which assumed a linear dependence between final height and BA). We decided that a quadratic relation was adequate for our purpose (in the Result section we mentioned that we have also tried fitting a third-order polynomial). Whether we judge the TW method using TW2, TW3, or the summed maturity score is unimportant, because these are related exactly by a non-linear transformation, which in a small interval of BA is well approximated by a second-order polynomial, so such a transformation is well accommodated by parabolic fits. It is only the BA method's ability to *order subjects of a given age correctly with respect to growth potential* which is assessed.

The residual error of the *gp* prediction models comes from three sources: (A) inaccuracy of the BA method, including rater variability and imperfect relation to the 'true' maturity; (B) measurement errors in height and adult height, and finally (C) some error comes from leaving out other factors than BA and age in the modeling of the growth potential, for instance body weight, time of menarche, and genetic and environmental factors in general. It should also be remembered that we are predicting the growth of the axial skeleton from the maturity of another part of skeleton, the hand, and these two parts may not mature in synchrony.

The point is that the contributions B and C are common to all BA methods, so the GPPEs for different BA methods are a measure of their accuracy or validity. This discussion also shows that BA methods must be compared on the *same* study data – different studies might yield different contributions from B and C.

Comparison of Manual TW and GP Ratings

The comparison between manual TW and GP rating showed that manual GP BA was significantly better. The TW ratings were used as a basis for the TW3 formulae for height prediction [1]. Hence, it seems that the errors in the TW3 adult height prediction model would have been reduced by using the GP BAs instead. To appreciate the size of this effect, one could take 3.20 cm as the typical prediction SD error found in Tanner et al. [1], and this would

then be reduced to 3.05 cm with GP BA ratings. As mentioned above, this improvement would be significant ($p < 0.05$).

The better performance of GP BA mainly stemmed from the conspicuous difference for boys at CA = 13 years, which is clearly seen in figure 2: TW3 BA is seen to be inaccurate in the BA range 12–13.5 years. This observation is supported by table 10 of Tanner et al. [1], in which the residual SD for boys showed a significantly larger error at CA = 13 years, corresponding to the peak in figure 4.

There were distinct differences in performance between boys and girls. Manual GP BA was much better than TW3 BA in boys, while they performed equally well in the girls. This can be explained by the mechanisms of the two rating methods: In the TW method the boys and girls are rated with the same nine stages that span the maturity range, whereas in the GP system there are 31 plates for the boys and 27 plates for the girls, so the GP system is more fine-grained for boys compared to girls, which could have led to the observed effect.

Figure 3 displays an interesting difference between GP and TW3 BA. The TW3 BA scale stops at 15 years for girls, while the GP method continues up to 18 years. Rating maturity all the way up to 18 leads to a lower *gp* prediction error (0.34 vs. 0.42).

Comparison with Automated Rating

The results for both sexes shown in table 1 indicated a slightly (but not significantly) better performance of BoneXpert (GPPE = 1.23) compared with manual GP rating (GPPE = 1.26). This is evidence that the automated method does indeed provide a BA determination which is as least as valid as the manual GP rating. Compared to manual GP ratings, BoneXpert GP ratings yielded a larger GPPE for boys and a smaller GPPE for girls. This may partly be due the above-mentioned fact that manual GP rating uses relatively few plates for the girls. This result presents a challenge for future versions of BoneXpert, which should be able to improve the BA rating of boys in particular.

Excluding the Wrist

BoneXpert's GP BA is formed as the average of the BA of the 13 RUS bones, i.e. radius and ulna each contribute 7.7% of the information. The BoneXpert GPshort BA method excludes the wrist (radius and ulna) from the average, but this hardly changed the GPPE (table 1). This shows that one does not need to include the wrist in BA rating as far as prediction of adult height is concerned. This finding may also shed further light on why TW3 rating was worse than GP rating: the TW3 method assigns an extraordinarily large weight of 40% to the wrist. Avoiding exposure of the wrist in BA X-rays would have the benefit of reducing radiation dose.

In Tanner et al. [9] it was shown that the carpals contribute no information to prediction of adult height. To our knowledge, there have never been any systematic studies of the relevance of including the wrist in BA assessment. Tanner always considered radius and ulna to be indispensable members of the RUS bones. This seems to be based on the intuitive judgement that: (a) one should include as many bones as possible to average out uncertainties in the rating of individual bones; (b) radius and ulna could be considered to be representative of the long bones in the body, in particular the femur, a major contributor to stature, and (c) in the RUS system, there is an appealing symmetry that radius, ulna, ray 1, ray 3, and ray 5 each have 20% weight.

There are other problems related to the wrist, because the pose of these two bones is more difficult to control, and the projection of the three-dimensional bone onto a plane under varying angles makes the interpretation difficult. The ulna is notoriously difficult to rate, manually as well as for BoneXpert. The lack of usefulness of the wrist found in this study could therefore also to some extent be due to BoneXpert's limited ability to analyze these bones.

Routine GP Ratings

It is likely that the manual BA ratings in the 1ZLS are more precise, i.e. have smaller rater variability than ratings performed in clinical practice, because these images were rated as an unblinded *series* rather than as independent images. Having the previous and following images in a series helps the rater to avoid outliers since he/she expects BA to progress steadily in healthy children. Present members of the Zurich group have confirmed that adjustments were made for these ratings to ensure steady progression [R. Largo and L. Molinari, pers. commun.]. This means that if a routine, clinical rater were to re-rate these X-rays in a random, blinded manner, he/she would incur larger rater errors and therefore larger GPPE. BoneXpert always rates images independently of each other, ignorant of identity, ethnicity, age, or medical history. Therefore the relative advantage of BoneXpert ratings observed in this study is likely to be larger in clinical practice.

Conclusion

Different BA rating systems differ from each other by SDs of at least 0.5 years, i.e. a 95% confidence interval of one type of BA rating relative to another is at least ±1 year. While a 1-year difference is certainly clinically relevant, we have hitherto been unable to assess which BA method is most valid. Studies have to date merely assessed the precision of various BA methods, i.e. their ability to obtain the same result when rating anew. By providing an objective reference based on prediction of adult height, the new framework presented here offers a means of establishing a golden standard in BA assessment. In the present study it was shown that the validity of BoneXpert GP BA was significantly better than that of manual TW3 BA in the 1ZLS.

Acknowledgement

Novo Nordisk is acknowledged for providing access to the film scanner. We thank Elisabeth Kaelin, Luciano Molinari and Jon Caflisch for data management of the 1ZLS.

Disclosure Statement

The first author is the owner of Visiana, which holds and markets the BoneXpert medical device for automated determination of BA.

References

1 Tanner JM, Healy M, Goldstein H, Cameron N: Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method), ed 3. London, Saunders/Harcourt, 2001.

2 Greulich WW, Pyle SI: Radiographic Atlas of Skeletal Development of the Hand and Wrist, ed 2. Stanford, Stanford University Press, 1959.

3 Thodberg HH, Kreiborg S, Juul A, Pedersen KD: The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2009;28:52–66.

4 Thodberg HH: An automatic method for determination of bone age. J Clin Endocrinol Metab 2009;94:2239–2244.

5 Bayley N, Pinneau SR: Tables for predicting adult height from skeletal age: revised for use with the Greulich-Pyle hand standards. J Pediatr 1952;40:423–441.

6 Martin DD, Deusch D, Schweizer R, Binder G, Thodberg HH, Ranke MB: Clinical application of automated Greulich-Pyle bone age in children with short stature, Pediatr Radiol 2009;39:598–607.

7 Van Rijn RR, Lequin MH, Thodberg HH: Automatic determination of Greulich and Pyle bone age in healthy Dutch children. Pediatr Radiol 2009;39:591–597.

8 Martin DD, Neuhof J, Jenni OG, Ranke MB, Thodberg HH: Automatic determination of left- and right-hand bone age in the First Zurich Longitudinal Study. Horm Res Paediatr 2010 (in press).

9 Tanner JM, Whitehouse RH, Marshall WA, Carter BS: Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4–16 with allowance for midparent height. Arch Dis Child 1975;50:14–26.

10 Roche AF, Wainer H, Thiessen D: The RWT method for the prediction of adult stature. Pediatrics 1975;56:1027–1033.

11 Tanner J, Landt K, Cameron N, Carter B, Patel J: Prediction of adult height from height and bone age in childhood. A new system of equations (TW Mark II) based on a sample including very tall and very short children. Arch Dis Child 1983;58:767–776.

12 Thodberg HH, Jenni OG, Caflisch J, Ranke MB, Martin DD: Prediction of adult height based on automated determination of bone age. J Clin Endocrinol Metab, 2009 epublication ahead of print, November 19, 2009, doi:10.1210/jc.2009-1429.